DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health

Division of Research Resources
Biotechnology Resources Program
Annual Progress Report
PART I

1. PHS GRANT NUMBER: P41RR01685-06

2. TITLE OF GRANT: BIONET, National Computer Resource for Molecular Biology

3. NAME OF RECIPIENT INSTITUTION: IntelliGenetics, Inc.

4. HEALTH PROFESSIONAL SCHOOL (If applicable):

5. REPORTING PERIOD:

5a. FROM (Month, Day, Year): 03-01-88

5b. TO (Month, Day, Year): 02-28-89

6. PRINCIPAL INVESTIGATOR:

6a. NAME: Dr. Michael J. Kelly

6b. TITLE: President, IntelliGenetics

6c. SIGNATURE: Michael J. Kelly

7. DATE SIGNED (Month, Day, Year): December 12, 1988

8. TELEPHONE (Include Area Code): 415-962-7313

2. Description of Program Activities

This section of our Annual Report provides statistical information on the use of the BIONET^{tm} Resource. The period covered is 12/87 - 11/88, to coincide with the dates of preparation of our Report and to follow our procedure of providing a full year's statistical information to compare with previous years' Reports.

Individual sections are prepared under guidelines discussed previously with BRTP staff and used in our previous Reports. We use a format for reporting the hundreds of individual Principal Investigator's use that is easy for us to generate while retaining the critical information necessary for BRTP in its internal and governmental reporting requirements. Complete research abstracts are kept at IntelliGenetics and are available upon request if needed.

The BIONET User community is divided into different classes, representing different levels of use of the computer system and staff resources, as follows:

- **Class I.** Class I users represent the Service component of the scientific community. They participate in the electronic communications facilities of BIONET (bulletin boards and electronic mail), and use the Core and Contributed Software libraries to pursue their research;
- **Class II.** Class II users represent the Collaborative component of the user community. Scientists in Class II enjoy all benefits of Class I use, and in addition contribute software and expertise to BIONET, working closely with BIONET staff. This category also includes bulletin board leaders, accounts for other related Resources (GenBank, NBRF/PIR, Dana Farber, etc.), and National Advisory Committee members.
- **Class III.** Class III accounts are for BIONET satellite communications. These accounts have the same system privileges as Class IV below but are provided free of charge.
- **Class IV.** Class IV users consist of those scientists who wish access only to the electronic communication facilities of BIONET. They are given access to the electronic mail and bulletin board facilities.
- **Class V.** Class V was implemented in response to a decision of our National Advisory Committee. Class V accounts are similar to Class IV but are for use by industrial scientists. Class V users sign an agreement stating that their communications account will be used only for scientific purposes and not for commercial advertising or other promotional purposes. Because of our ARPANET connection restrictions, users in this class do not have access to the TELNET and FTP programs which allow direct contact to other computers on the ARPANET.

Information on the number of PI's by Class is summarized in Table 2-1.

Table 2-1: Summary of the BIONET User Community

Class I	813
Class II	26
Class III	7
Class IV	19
Class V	2
Total	867

The total number of laboratories with access to BIONET, 867, is a 31 percent increase over the total of 660 presented in our last annual report! This clearly demonstrates the continuing demand for and the quality of the BIONET service. Besides having a small rate of discontinued accounts (9.6%, below) the resource has continued to grow at a remarkable rate. Between 12/87 - 11/88, 270 new labs opened accounts on BIONET, and 63 (9.6%) labs discontinued their subscriptions.

The current number of BIONET PI's (867) represents about 14% of all the NIH extramural investigators (some 6100 grants total). Actually, the total number of NIH-funded investigators may be significantly less than 6100 since many investigators hold more than one grant. Since only about half of the NIH grantees mention DNA sequence, cloning, or recombinant DNA in their grants, we would estimate that the 14% of NIH grantees that are on BIONET represents between 25 and 30% of all the scientists who could possibly make use of the BIONET facility. BIONET has considerable potential for further growth. A major factor in this continued demand is the increasing recognition by the scientific community of the need for a facility which provides rapid access to molecular biology databases and which also can serve as a hub for electronic communications. As the amount of sequence data continues to expand, the central role of BIONET will increase in value.

Considerable time and effort has been expended in developing the BIONET resource to its current state. We trust that the NIH recognizes the value of this resource to the community and is aware of the dislocation that would occur in a significant number of important research projects if any disruption occurred in the service.

2.1 Scientific Subprojects

2.1.1 Collaborative Research and Service

In the following section we report the use of the BIONET Resource for Class I-V users. The "Usage Factor" is reported as both central processor unit (cpu) time, in minutes, and connect time, in hours, for each Principal Investigator on the BIONET DEC 2065 computer. These values are the sum of all usage by the PI and his or her group members ("Sub-I's").

We note that, starting in August 1988, BIONET made available a Sun 3/280 computer to handle database searches. This was necessary because of the tremendous user demands being made on the DEC computer. Since user demand was affecting the response time on the latter system, we made access to the Sun available despite its lack of accounting software. Actual user totals may therefore be higher in many cases than listed below. We are currently working on implementing accounting software on the Sun systems and statistics on its use will be available in the near future.

We report data only on those PI groups that have used the Resource during the past 12 months. Of the 867 groups with active BIONET accounts, 693, representing about 2430 individual investigators, have accessed BIONET. Last year's access total was 530 laboratories. Thus the number of groups who have accessed the system increased by 31% as compared to last year.

If frequent users are defined as those laboratories utilizing 60 or more connect hours or 120 or more cpu minutes during the year, there were 407 groups in the "frequent user" category this year (47% of all current accounts). This represents an increase of 41% over this same category last year (288).

There are 174 accounts on the system that were inactive during the year. Most of the accounts in this category were inactive either because they were created near the end of our accounting year and not yet utilized, or because they fell into the category of complimentary (mainly foreign) accounts. Foreign users are not charged an access fee but must pay their own telecommunications charges. These accounts (now totaling 111) have been used infrequently because of the expense of international telecommunications access. However, since they utilize little of our resources, the accounts have been maintained on the system and have accumulated over the past five years.

The summary usage statistics for each laboratory group follow below. Detailed usage statistics for each individual user are maintained by the BIONET computer and are available to interested parties.

We do not report Resource staff hours nor BRTP funds allocated for individual PI's because it is impossible to allocate these rationally to such a large user community. Summary information on allocation of staff hours is given in the *Resource Summary Table*.

2.1.2 Technological Research and Development

We report on the standard form summary information for our Technological Research projects.

In past years the Resource Technology used was the DEC-2065 computer. This year most work by the staff has been performed on BIONET's new Sun computers, and accounting software for reporting CPU minutes is not currently available on those machines.

The Usage Factor was previously reported as minutes of cpu time used for the project, but is not provided this year due to the lack of usage statistics on the Sun computers.

Resource Staff Hours are based on time estimates of work reported on each project. For use in further calculations (below), hours listed for each individual in the table are divided by total annual hours per individual. This yields a fractional time or "FT" for each individual on each project.

"BRTP Funds Allocated" are calculated as the sum of the following components:

- **Actual Personnel Costs.** The personnel costs for each project are derived by multiplying the above FT for each BIONET staff person's time spent on the project times their respective annual salary plus fringe benefits; the actual personnel cost is the sum of these individual figures.
- **Consultant Costs.** The FT spent by a consultant involved in a project is multiplied times the total consulting cost for the consultant; these are summed for each project where appropriate.
- **Fraction of Awarded Funds.** The fraction of total awarded funds for each project is derived by multiplying the **fractional time** spent on each project (defined below) by the **awarded funds** (defined below). The **fractional time** is determined from the sum of hours spent on the project by the investigators listed divided by the sum of hours spent on BIONET by all investigators. For this calculation we have used the actual time spent from 12/87 through 11/88. Although this period is three months out of phase with the actual grant period, we do not think the fractional time spent will change significantly during the next three months of the current period. In computation of **awarded funds**, we include only the funds allocated in the grant categories of *Supplies*, *Travel*, and *Other Expenses*. The categories of *Personnel* and *Consultants* are accounted for in the previous two computations and the remaining grant category *Equipment* is accounted for in the *Resource Summary Table*.
- **Indirect Costs.** Indirect costs are allocated by multiplying the total awarded indirect costs for this reporting period by the fractional time devoted to each project. The fractional time is determined from the sum of hours spent on the project by all investigators divided by the sum of hours spent on BIONET by all investigators.

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A

INSTITUTION: IntelliGenetics

GRANT NUMBER P41RR01685-06

PERIOD: March 1, 1988 to February 28, 1989

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

[X] TECHNOLOGICAL RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH & SERVICE [ ] TRAINING

1 Descriptive Title (80 characters) Abstract	2 Science Code Axis I	2 Science Code Axis II	3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.	4 USAGE FACTOR Resource Technology a/	4 USAGE FACTOR Hours Used b/	4 USAGE FACTOR Resource Staff Hours	5 BRTP Funds Allocated (direct and indirect)
Comparative Sequence Analysis and Software Development. Study and classification of human Alu and Kpnl sequences. Development of multiple sequence alignment editor.	9	42,58	a,b Jurka, Jerzy W. Ph.D. Horng, Liang J. M.S. Maulik, Sunil Ph.D. c IntelliGenetics	Sun 3/280 Sun3/60's	N/A N/A	1246 876 7	183,603
Protein Secondary Structure Anaylsis. Use of comparative sequence analysis and 3-D protein structural information to investigate determinants of protein secondary structure.	9	42	a,b Jurka, Jerzy W. Ph.D. Maulik, Sunil Ph.D. c IntelliGenetics	Sun 3/280 Sun3/60's	N/A N/A	560 155	64,224
RICH Program Development . Algorithm development and coding of software to search sequence databases for sequences of defined composition.	9	42	a,b Maulik, Sunil Ph.D. c IntelliGenetics	Sun 3/280 Sun3/60's	N/A N/A	556	48,685
CUMULATIVE TOTALS: No. subprojects							

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions ...

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A **INSTITUTION:** IntelliGenetics, Inc.

GRANT NUMBER P41RR01685-06 **PERIOD:** March 1, 1988 to February 28, 1989

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

[X] TECHNOLOGICAL RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH & SERVICE [ ] TRAINING

1 Descriptive Title (80 characters) Abstract	2 Science Code Axis I	2 Science Code Axis II	3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.	4 USAGE FACTOR Resource Technology a/	4 USAGE FACTOR Hours Used b/	4 USAGE FACTOR Resource Staff Hours	5 BRTP Funds Allocated (direct and indirect)
User Interface Development. Development of a prototype BIONET user interface based on Hypercard software for the Apple Macintosh computer.	9	42	a,b Maulik, Sunil Ph.D. c IntelliGenetics	Macintosh	N/A N/A	318	27,845
Network Mail Server for Sequence Database Searches. Development of FASTA-MAIL software for remote data-base searching by electronic mail.	9	42	a,b Dautricourt, J.P. Ph.D. Lear, Eliot B.S. Liebschutz, Robert B.A. Yeh, Spencer B.S. c IntelliGenetics	Sun 3/280 Sun3/60's DEC 2065	N/A N/A N/A	67 233 20 30	22,627
Central Resource Development Hardware configuration and testing for Sun Microsystems Network. Systems software development.	9	42	a,b Lear, Eliot B.S. Liebschutz, Robert B.A. Diaz, Ron B.S.	Sun 3/280 Sun3/60's	N/A N/A	307 367 81	66,615
CUMULATIVE TOTALS: No. subprojects							

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions ...

7a

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A — INSTITUTION: IntelliGenetics, Inc.

GRANT NUMBER P41RR01685-06 — PERIOD: March 1, 1988 to February 28, 1989

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

[X] TECHNOLOGICAL RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH & SERVICE [ ] TRAINING

1 Descriptive Title (80 characters) / Abstract	2 Science Code Axis I	2 Science Code Axis II	3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.	4 USAGE FACTOR: Resource Technology a/	4 USAGE FACTOR: Hours Used b/	4 USAGE FACTOR: Resource Staff Hours	5 BRTP Funds Allocated (direct and indirect)
Elecronic Communication. Establishment of communication links between BIONET and remote computer sites. Use of ARPANET, USENET, and BITNET for networked sites. Development of communications software for mail and bulletin boards on UNIX and VAX/VMS sites. Development of international newsgroup distribution network for biologists.	9	40,42	a,b Diaz, Ron B.S. Lear, Eliot B.S. Kristofferson, David Ph.D.	Sun 3/280 Sun3/60's MicroVAX DEC 2065	N/A	456 243 200	78,580
Data Submission software. Addition of user requested features to BIONET XGENPUB data submission program.	9	42	a,b Kanerva, Lauri	DEC 2065	N/A	45	3,679
CUMULATIVE TOTALS: No. subprojects	8					5767	495,858

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions ...

2.1.3 Training

We report summary information for our Training program. The sites at which BIONET provided some level of training are named here and are discussed in more detail in the *Narrative Description* section.

The method for calculation of Usage Factors and BRTP Funds Allocated is the same as that described above under Technological Research and Development.

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A **INSTITUTION:** IntelliGenetics

GRANT NUMBER P41RR01685-06 **PERIOD:** March 1, 1988 to February 28, 1989

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

[ ] TECHNOLOGICAL RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH & SERVICE [X] TRAINING

1 Descriptive Title (80 characters) Abstract	2 Science Code Axis I	2 Science Code Axis II	3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.	4 USAGE FACTOR Resource Technology a/	4 USAGE FACTOR Hours Used b/	4 USAGE FACTOR Resource Staff Hours	5 BRTP Funds Allocated (direct and indirect)
BIONET Training Program Support of training for BIONET scientist including four in-house training sessions at IntelliGenetics, phone trainings, preparation of new training documentation, and outside demonstrations at Rutgers, the NIH, FASEB, Windsor, Ontario, and Ohio State at Wooster.	9	40,68	a,b Johncox, Vickie B.S. Bigham, Nancy M.S. Berg, Kathryn M.P.A. Davis, Karen Ph.D. Kristofferson, David Ph.D. Maulik, Sunil Ph.D. Yeh, Spencer B.S. Benton-Vosman, Trish M.S. Swank, Beth B.S. c IntelliGenetics	DEC-2065 Sun 3/280	N/A " " " " " " " "	68 3 8 99 59 181 61 9 6	41,799
CUMULATIVE TOTALS: No. subprojects	1					494	41,799

a/ Identify Resource Technologies Used. b/ Give Hours Resource Technologies Used. See Instructions ...

2.2 Books, Papers, Abstracts

We list on the next BRTP form the following research reports which detail some of BIONET's progress during this last year. Copies of these reports are available in *Appendix I.*

The second form lists papers that have trained or informed scientists about the resource. Copies of these papers are in *Appendix II.*

PART II, SECTION B	GRANT NUMBER	P	4	1	R	R	0	1	6	8	5	-	0	6

INSTITUTION: IntelliGenetics **REPORT PERIOD:** 3/1/88 **to** 2/28/89

Fill out a separate form for each of the following categories: Check one.

[X] TECHNOLOGICAL RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH [ ] TRAINING

Author(s) Title of Article, Journal, Vol., Number Pages(e.g., 44-48), Year Published.

Jurka, J. and T. Smith. (1988). Proc. Natl. Acad. Sci. USA. 85, 4775-4778. A fundamental division in the Alu family of repeated sequences.

Faulkner, D. V. and J. Jurka. (1988). Trends Biochem. Sci. 13, 321-322. Multiple aligned sequence editor.

Jurka, J.,T. F. Smith, and D. Labuda. (1988). Nucl. Acids Res. 16, 766. Small cytoplasmic Ro RNA pseudogene and an Alu repeat in the human α-1 globingene.

Jurka, J. and R. J. Britten. (1988). Cold Spr. Harb. Symp. abstr. Evolution of human Alu repeats: implications for genome studies.

Holsztynska, E., D. J. Waxman, and J. Jurka. (1988). Protein Society Mtg. abstr. Studies on rat liver cytochromes P450 using comparative sequence analysis.

Maulik, S. (1988). Protein Society Mtg. abstr. Locating amino acid patterns in proteins by composition.

Cumulative Number Published:		**Books** --	**Papers** 3	**Abstracts** 3
Cumulative Number in Press:		**Books** --	**Papers** --	**Abstracts** --

PART II, SECTION B	GRANT NUMBER	P	4	1	R	R	0	1	6	8	5	–	0	6

INSTITUTION: IntelliGenetics REPORT PERIOD: 3/1/88 to 2/28/89

Fill out a separate form for each of the following categories: Check one.

[ ] TECHNOLOGICAL RESEARCH & DEVELOPMENT [ ] COLLABORATIVE RESEARCH [X] TRAINING

Author(s) Title of Article, Journal, Vol., Number Pages(e.g., 44-48), Year Published.

Maulik, S. (in press). Protein Sequence and Data Analysis. Protein Databases and Software on BIONET.

Cumulative Number Published:		Books --	Papers --	Abstracts --
Cumulative Number in Press:		Books --	Papers 1	Abstracts --

We report the publications by members of the BIONET scientific community on a version of the special form provided by BRTP. These publications have **ALL** arisen from use of BIONET, although support by BIONET and the NIH has not always been acknowledged.

The figures on *Cumulative Number Published* refer to the current year alone. We have received 44 papers that were published or are in press. The total for the three previous years was 250, bringing the overall total to 294. We note that the actual number of publications which involved the use of BIONET is undoubtedly higher because many investigators have not yet replied to our requests for reprints, and the requirement to acknowledge BIONET is not strictly followed.

PART II, SECTION B	GRANT NUMBER	P	4	1	R	R	0	1	6	8	5	–	0	6

INSTITUTION: IntelliGenetics **REPORT PERIOD:** 3/1/88 to 2/28/89

Fill out a separate form for each of the following categories: Check one.

[ ] TECHNOLOGICAL RESEARCH & DEVELOPMENT [X] COLLABORATIVE RESEARCH [ ] TRAINING

Author(s) **Title of Article, Journal, Vol., Number Pages(e.g., 44-48), Year Published.**

Please see the following pages.

Cumulative Number Published:	42	**Books** --	**Papers** 41	**Abstracts** 1
Cumulative Number in Press:	2	**Books** --	**Papers** 2	**Abstracts** --

References

Akella, R., P. Arasu, and A. B. Vaidya. (1988) Molecular and Biochemical Parasitology, Vol. 30, pp. 165-174. "Molecular clones of @-tubulin genes of Plasmodium yoelii reveal an unusual feature of the carboxy terminus".

Allison, L.A. , J. K.-C. Wong, V. D. Fitzpatrick, M. Moyle, and J. Ingels. (1988) Molecular and Cellular Biology, Vol. 8, pp. 321-329. "The C-Terminal Domain of the Largest Subunit of RNA Polymerase II of Saccharomyces cerevisiae, Drosophila melanogaster, and Mammal: a Conserved Structure with an Essential Function".

Auger, I. E. , and C. E. Lawrence. (1988) Society of Mathematical Biology, Vol. 0092-8240, pp 1-16. " Algorithms For The Optimal Identification of Segment Neighborhoods".

Batter, D. K. , S. R. D'Mello, L. M. Turzai, H. B. Hughes III, A. E. Gioio, and B. B. Kaplan. (1988) The Journal of Neurosciences Research, Vol. 19, pp 367-376. " The Complete Nucleotide Sequence and Structure of the Gene Encoding Bovine Phenylethanolamine N-Methyltransferase".

Bray, S. J. , W. A. Johnson, J. Hirsh, U. Heberlein, and R. Tijian. (1988) The EMBO Journal, Vol. 7, pp. 177-188. "A cis-acting element and associated binding factor required for CNS expression of the Drosophila melanogaster dopa decarboxylase gene".

Brayton, K. A., J. Amim, H. Qui, R. Yazdanparast, M. A. Ghatei, J. M. Polak, S. R. Bloom, and J. E. Dixon. (submitted) DNA. "Cloning, Characterization, and Sequence of a Porcine cDNA Encoding a Novel Secreted Neuronal and Endocrine Protein".

Burns, G., T. Brown, K. Hatter, J . R. Sokatch. (1988) Eur. J. Biochem., Vol. 176, pp. 165-169. "Comparison of the amino acid sequences of the transacylase components of branched chain oxoacid dehydrogenase of Pseudomonas putida, and the pyruvate and 2-oxoglutarate dehydrogenases of Escherichia coli".

Burns, G. , T. Brown, K. Hatter, J. M. Idriss, and J. R. Sokatch. (1988) Eur. J. Biochem., Vol. 176, pp. 311-317. "Similarity of the E1 subunits of branched-chain-oxoacid dehydrogenase from Pseudomonas putida to the corresponding subunits of mammalian branched-chain-oxoacid and pyruvate dehydrogenases".

Burns Jr. , J. M. , T. M. Daly, A. B. Vaidya, and C. A. Long. (1988) Proc. Natl. Acad. Sci. USA, Vol. 85, pp. 602-606. "The 3' portion of the gene for a Plasmodium yoelii merozoite surface antigen encodes the epitope recognized by a protective monoclonal antibody".

Cao, M., X. Xiao, B. Egbert, T. M. Darragh, and T. S. B. Yen. "Rapid Detection of Cutaneous Herpes Simplex Virus Infection with the Polymerase Chain Reaction". (in press 9/30/88)

Chang, J. H., C. Tamba, S. Dumbar, and M. O. J. Olson. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 12824-12827. "cDNA and Deduced Primary Structure of Rat Protein B23, a Nucleolar Protein Containing Highly Conserved Sequences*".

Cohen, J. I. , R. H. Miller, B. Rosenblum, K. Denniston, J. L. Gerin, and R. H. Purcell. (1988) Virology, Vol. 162, pp. 12-20. :"Sequence Comparison of Woodchuck Hepatitis Virus Replicative Forms Shows Conservation of the Geneome".

Cooke, N. E., J. Ray, J. G. Emery, and S. A. Liebhaber. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 9001-9006. "Two Distinct Species of Human Growth Hormone-variant mRNA in the Human Placenta Predict the Expression of Novel Growth Hormone Proteins".

Dickey, L. F. , E. C. Theil, Y. H. Wang, G. E. Shulls, and I. A. Wortman III. (1988) The Journal of Biological Chemistry, Vol. 263, pp 3071-3074. "The Importance of the 3' - Untranslated Region in the Translational Control of Ferritin mRNA*"

D'Mello , S. R. , E. P. Weisberg, M. K. Stachowiak, L. M. Turzai, A. E. Gioio, and B. B. Kaplan. (1988) The Journal of Neurosciences Research, Vol. 19, pp. 440-449. "Isolation and Nucleotide Sequence of a cDNA Clone Encoding Bovine Adrenal Tyrosine Hydreoxylase: Comparative Analysis of Tyrosine Hydroxylase Gene Products".

Jeppesen, C., B. Stebbins-Boaz, and S. A. Gerbi. (1988) Nucleic Acids Research, Vol. 16, pp. 2127-2148. "Nucleotide sequence determination and secondary structure of Xenopus U3 snRNA".

Kokubu, F., K. Hinds, R. Litman, M. J. Shamblott, and G. W. Litman. (1988) The EMBO Journal, Vol. 7, pp. 1979-1988. "Complete structure and organization of immunoglobulin heavy chain constant region genes in a phylogenetically primitive vertebrate".

Kokubu, F., R. Litman, M. J. Shamblott, K. Hinds, and G. W. Litman. (1988) The EMBO Journal, Vol. 7, pp. 3413-3422. "Diverse organization of immunoglobulin VH gene loci in a primitive vertebrate".

Krawetz, S. A., W. Connor, and G. H. Dixon. (1988) DNA, Vol 6, pp. 47-57. "Cloning of Bovine P1 Protamine cDNA and the Evolution of Vertebrate P1 Protamines".

Krawetz, S. A., W. Connor, and G. H. Dixon. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 321-326. "Bovine Protamine Genes Cantain a Single Intron".

Krawetz, S. A., and G. A. Dixon. (1988) Journal of Molecular Evolution, Vol. 27, pp. 291-297. "Sequence Similarities of the Protamine Genes: Implications for Regulation and Evolution".

Linskens, M. H. , and J. A. Huberman. (1988) Molecular and Cellular Biology, Vol. 8, pp 4927-4935. " Organization of Replication of Ribosomal DNA in Saccharomyces cerevisiae".

Manly, K. F. , G. R. Anderson, and D. L. Stoler. (1988) Journal of Virology, Vol. 62, pp. 3540-3543. " Harvey Sarcoma Virus Genome Contains No Extensive Sequences Unrelated to Those of Other Retroviruses except *ras*".

Mayfield, J. E., B. J. Bricker, H. Godfrey, R. M. Crosby, D. J. Knight, S. M. Halling, D. Balinsky, and L. B. Tabatabai. (1988) Gene, vol 63, pp 1-9. "The cloning, expression, and nucleotide sequence of a gene coding for an immunogenic Brucella abortus protein."

Miller, R. H. (1988) Science, Vol. 239, pp. 1420-1422. "Human Immunodeficiency Virus May Encode a Novel Protein on the Genomic DNA Plus Strand".

Miller, R. H. (1988) Virology, Vol. 164, pp. 147-155. "Close Evolutionary Relatedness of the Hepatitis B Virus and Murine Leukemia Virus Polymerase Gene Sequence".

Nagle, G. T., S. D. Painter, J. E. Blankenship, and A. Kurosky. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 9223-9237. "Proteolytic Processing of Egg-laying Hormone-related Precursors in *Aplysia*".

Nagle, G. T. , S. D. Painter, J. E. Blankenship, J. V. A. Choate, and A. Kurosky. (1988) Peptides, Vol. 9, pp. 867-872. "The Bag Cell Egg-Laying Hormones of *Aplysia brasiliana* and *Aplysia californica* are Indentical".

Nees, D. W. , P. A. Stein, and R. A. Ludwig. (1988) Nucleic Acids Research, Vol. 16, pp. 9839-9853. "*The Azorhizobium caulinodans nifA* gene: dentification of upstream-activation sequences including a new element, the 'anaerobo' "

Oka, Y., and C. A. Thomas, Jr. . (1988) Nucleic Acids Research, Vol. 15, pg. 8877-8898. "The cohering telomeres of Oxytricha".

Rimsky, L., J. Hauber, M. Dukovich, M. H. Malim, A. Langlois, B. R. Cullen, and W. C. Greene. (1988) Nature, Vol. 335, pp.738-740. "Functional replacement of the HIV-1 rev protein by the HTLV-1 rex protein".

Singh, S. V., H. Ahmad, A. Kurosky, and Y. C. Awasthi. (1988) Archives of Biochemistry and Biophysics, Vol. 264, pp. 13-22. "Purification and Characterization of Unique Glutathione S-Transferases from Human Muscle".

Suplick, K., R. Akella, A. Saul, and A. B. Vaidya. (1988) Molecular and Biochemical Parasitology, Vol. 30, pp. 289-290. "Molecular cloning and partial sequence of a 5.8 kilobase pair repetitive DNA from *Plasmodium falcparum*".

Sung, S. J., J. M. Bjorndahl, C. Y. Wang, H. T. Koa, and S. M. Fu. (1988) J. Exp. Med., Vol. 167, pp. 937-953. "Production of Tumor Necrosis Factor/Cachectin by Human T Cell Lines and Peripheral Blood T Lymphocytes Stimulated By Phorbol Myristate Acetate and Anti-CD3 Antibody".

Upton, C., J. L. Macen, R. A. Maranchuk, A. M. DeLange, and G. McFadden. (1988) Virology, Vol. 0042-6822 , pp. 229-239. Tumorigenic Poxviruses: Fine Analysis of the Recombination Junction in Malignant Rabbit Fibroma Virus, a Recombinant between Shope Fibroma Virus and Myxoma Virus".

Vodkin, M. D. , and J. C. Williams. (1988) Journal of Bacteriology, Mar. 88, pp. 1227-1234. " A Heat Shock Operon in Coxiella burnetii Produces a Major Antigen Homologous to a Protein in Both Mycobacteria and Escherichia coli"

Vold, B. S. , C. J. Green, N. Narasimhan, M. Strem, and J. N. Hansen. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 14485-14490. "Transcriptional Analysis of Bacillus subtilis rRNA-tRNA Operons".

Wang, Y. H. , S. R. Sczekan, and E. C. Theil. (1988) Metal Ion Homeostasis: Molecular Biology and Chemistry. UCLA Symposia on Molecular and Cellular Biology. Vol. 98, D. Winge & D. Hamer, Editors, Alan R. Liss, Inc., New York.

Weber, J. L. . (1988) Molecular and Biochemical Parasitology, Vol. 29, pp. 117-124. "Interspersed repetitive DNA from *Plasmodium falciparum*".

Weber, J. L., J. A. Lyon, R. H. Wolff, T. Hall, G. H. Lowell, and J. D. Chulay. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 11421-11425. "Primary Structure of a *Plasmodium falciparum* Malaria Antigen Located at the Merozoite Surface and within the Parasitophorous Vacuole*".

Westaway, S. K., E. M. Phizicky, and J. Ableson. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 3171-3176. " Structure and Function of the Yeast tRNA Ligage Gene*".

Winkfien, R. J., R. D. Moir, S. A. Krawetz, J. Blanco, J. C. States, and G. H. Dixon. (1988) Eur. J. Biochem., Vol. 176, pp. 255-264. "A New family of repetitive, retroposon-like sequences in the genome of the rainbow trout".

Wohlrab, H., R. T. Bronson, R. C. Lu, and V. Nameth. (1988) Biomedical and Biophysical Research Communications, Vol. 154, pp. 1130-1136. "Towards a Biomarker of Mammalian Senescence: Carbonic Anhydrase III".

Xiao, X., M. Cao, T. R. Miller, Z. Y. Cao, and T. S. B. Yen. The Lancet, October 15, 1988 , p. 902. Papillomavirus DNA in Cervical Carcinoma Specimens from Central China. (abstract)

2.3 Resource Summary Table

The Resource Summary Table includes the totals from the previous sections of *Technological Research and Development* and *Training*. The totals for *Collaborative Research and Service* are derived from the following components:

- **Usage Factor**. The *CPU Min. Used* is the total for BIONET CPU use found in Table III-7 since this year cpu use on the DEC was essentially entirely consumed by the user community. Technological Research projects were performed on other hardware. The total of staff hours is obtained from our accounting totals of hours spent on BIONET projects and the value for Collaborative Research and Service is likewise the remainder after subtracting time spent in other categories.
- **BRTP Funds Allocated**. This is computed as the difference between the total budget for BIONET minus the categories of Technological R+D and Training, and minus the capital equipment expenditures for the year ($36,746) listed under Administration/Miscellaneous.

The category of *Collaborative Research and Service* includes an entry of $169,182 in the column *Other Funds*. This is the total money invoiced over the period 12/87 - 11/88 for subscription fees ($252,932) minus outstanding receivables of $83,750. Each PI is asked to pay an access fee to help defray the telecommunication costs for access to BIONET; this fee is currently $400/year. By agreement with BRTP, these access fees are not grant related income.

The balance of these fees carried forward from the previous year (as of 12/1/86) was $156,523. After twelve additional months of collecting subscription fees ($169,182 above) and disbursing them for telecommunication expenses ($244,321), the balance is now $81,384.

No facility staff computer time, work hours, or BRTP funds are allocated to the category of *Administration / Miscellaneous*; we consider such time and funds to be an integral part of the support of the other components of the Resource. We do include as Funds Allocated the $36,746 used to purchase items of capital equipment.

The category of down time includes the sum of scheduled and unscheduled maintenance on the DEC-2065 computer. In the period 12/87 - 11/88, there was a total of 149 hours (8924 cpu minutes) of downtime:

- 2545 cpu minutes of scheduled downtime for preventive maintenance and several system-related tasks.
- 6379 minutes of downtime were due to unscheduled maintenance.

The downtime reported in the Summary Table (8032 cpu min) is 90% of the total, reflecting BIONET's allocation of 90% of the machine. Note that the **total** unscheduled maintenance of 6379 cpu minutes is only 1.2% of the total cpu time available. Considering both categories of downtime, the machine has been available for use by BIONET scientists, 98.3% of the time, 24 hours a day, seven days a week. No funds have been allocated to this category.

PART II SECTION C RESOURCE SUMMARY TABLE

GRANT NUMBER P41RR01685-06 REPORT PERIOD 3/1/88 to 2/28/89

RESOURCE COMPONENT	Number Subproject	Number Publications	Number Investigators	USAGE FACTOR Resource a/ Technology	USAGE FACTOR CPU/ mins. used	USAGE FACTOR Resource Staff Hrs	BRP Funds Allocated $	Resource Fees $ Collected	Other Funds $
TECHNOLOGICAL RESEARCH & DEVELOPMENT	8	6	10	DEC 2065 Sun 3/280 Sun 3/60 MicroVax	N/A	5,767	495,858		
COLLABORATIVE RESEARCH & SERVICE	693	44	2,430	DEC 2065 Sun 3/280	420,932 N/A	11,103	962,651		169,182
TRAINING	1	1	9	DEC 2065 Sun 3/280	N/A	494	41,799		
ADMINISTRATION/ MISCELLANEOUS	see preface to this section.						36,746		
DOWN TIME					8,032				
GRAND TOTALS	702	51	2,449		N/A	17,364	1,537,054		169,182

3. Narrative Description

3.1 Summary of Research Progress

Although a very signficant fraction of staff time during our fifth year was involved with planning, grant writing, and other obligations for our renewal, important progress was still made on the Resource. The following sections describe in detail our accomplishments in the several components of the BIONET Resource. Here, in brief, are some of the most notable.

- The new computer network donated by Sun Microsystems is being prepared for direct access by the user community by the BIONET systems staff (Mr. Rob Liebschutz and Mr. Eliot Lear). As of August 1988 BIONET released the FASTA-MAIL program. This provided our users on the DEC with electronic mail access to high-speed database searches on our Sun 3/280 computer. Database search times were reduced from hours to tens of minutes or less, and the average mid-day user load on the DEC 2065 dropped by a factor of about three to five since it was no longer being used for these compute-intensive tasks. FASTA-MAIL used the FASTA program obtained from Dr. William Pearson, and the mail server portion was developed by Mr. Liebschutz, Mr. Lear, and Mr. Spencer Yeh.

- Dr. Jerzy Jurka, the BIONET Scientist, has published important work in the area of repetitive DNA sequence analysis. In conjunction with this research, new functionality has been added to the Multiple Aligned Sequence Editor (MASE) by the BIONET applications programmer, Mr. Liang Jen Horng. This editor was originally developed by Dr. Jurka and Donald Faulkner at Dana Farber's Molecular Biology Computer Research Resource and then extended at BIONET over the past year. Work on the editor has also involved collaborators from the machine learning group at the University of California at Santa Cruz.

- Dr. Sunil Maulik, BIONET's Senior Scientific Consultant, has continued work on the RICH program which performs database searches for sequences of defined percent composition. Dr. Maulik has obtained some interesting preliminary results with the program. He has also been involved in developing a new Hypercard ^{tm} -based user interface for BIONET.

- The electronic communications network was significantly enhanced. The efforts of Dr. David Kristofferson led to the formation of the international BIOSCI bulletin board network. Because scientists work on a variety of computer networks around the world, we recognized the necessity of developing a mechanism to allow all of them to communicate without the necessity of learning the peculiarities of accessing each network. We sought out computer sites on all major international networks and arranged to have parallel copies of the original BIONET bulletin boards accessible from these sites. Besides BIONET in the U.S., other major BIOSCI distribution sites are situated in England, Ireland, Sweden, and Finland. Recipients of the bulletin boards from these sites are located around the world from New Zealand and Australia, the Far East, and Israel, throughout Europe, and back to North America. The bulletin boards are available to users on the ARPANET, BITNET, EARN, Usenet, NSFnet, and JANET. Users in any particular location need only post or receive messages from their closest site. Any postings at any one site are automatically forwarded by the central BIOSCI sites to all other participants on all of the above-listed networks.

- Finally, the research conducted by BIONET's 867 laboratory groups was made significantly easier by a total revision of the BIONET documentation and the production of a new User Manual. This involved major efforts by BIONET staffer's Ms. Vickie Johncox, Mr. Spencer Yeh, and Ms. Kathryn Berg. The documentation was sent free of charge to all users on the system this past summer.

3.1.1 Service

The Service component of BIONET includes primarily Class I investigators who use the BIONET Core and Contributed program Libraries to support their research. BIONET user classes were explained above under *Description of Program Activities.* The BIONET consultants also provide support as needed by the other classes of BIONET users, but these groups are much smaller than the predominant class I group.

The computing assistance given by the staff can range from simply answering routine questions to providing sophisticated help on sequence alignments or complex database searches. Especially in the latter case the staff can make a significant contribution to the attainment of a BIONET user's research goals. Viewed in this light the distinction between "service" and "research" may be hard to discern.

Before going into the details of how the BIONET staff serves the user community we wish to provide several examples of how BIONET Class I investigators utilize the resource. These "case studies" will demonstrate the importance of the work being performed by the investigators which the BIONET staff serves.

3.1.1.1 Scientific Case Studies Using BIONET

The success of BIONET can be measured in several ways. For example, one can count the number of participating scientists, or count their publications. These numbers are interesting and impressive, but do not convey the high quality of work that is being done. It is very difficult to measure this quality objectively. We have examined the publications submitted to us, and have selected two that we feel represent some of the quality research done on the system. We present these as "case studies."

"Harvey sarcoma virus genome contains no extensive sequences unrelated to those of other retroviruses except *ras*." Kenneth F. Manly, Garth R. Anderson, and Daniel L. Stoler. ***Journal of Virology 62*****: 3540-3543 (1988).**

Dr. Manly's laboratory has been studying the VL30 multigene elements present in rats and mice with structures similar to those of retroviruses' and retrotransposons' genomes. Transcripts of these elements are packaged as pseudotypes by type C retroviruses, and, when introduced into cells, pseudotyped VL30 RNA can lead to integration at new sites in the cell genome, suggesting that they are an entirely new class of transposable elements. Further, both the Kirsten and Harvey murine sarcoma viruses are acute transforming viruses whose genomes comprise two distinct genomic elements; a *ras* oncogene and a VL30 sequence, both of which appear to contribute directly to oncogenic activity. VL30 element transcription is strongly induced as a cellular response to anoxic stress, and studies in Dr. Manly's laboratory had previously suggested that the rat VL30 sequences incorporated into Kirsten sarcoma virus genome might directly encode the major anoxic stress protein p34, lactate dehydrogenase k.

In order to characterise the VL30 domain from Harvey sarcoma virus (HaSV) and to evaluate its similarity to other retroviral sequences, Dr. Manly and his collaborators compared translated HaSV sequences with the entire NBRF-PIR database available on BIONET using the IFIND and XFASTP database similarity searching software. Translations were done in all three reading-frames using

the PEP program. Selected sequences were evaluated for significance with the XRDF program, which compares the observed similarity score with a group of similarity scores obtained by randomizing one of the sequences many times. Additionally, HaSV RNA subsequences which correspond to the regions of peptide similarity were searched against the GenBank viral nucleic acid sequences using the XFASTN program on BIONET. Terminator codons in the HaSV RNA sequence were converted to X's to allow their acceptance by XFASTP. (The X character is treated by XFASTP as an unknown residue and is given an intermediate relatedness score).

The results of the searches yielded eight major regions of sequence similarity with (optimized) similarity scores ranging from 47 to 428 and z-scores (the alignment score for the sequence expressed as the number of standard deviations above the mean of a set of scores from the randomized sequences) ranging from 5 to 42. (Table I, pg. 3541). Three regions showed greatest similarity with *gag* sequences of feline sarcoma virus, with the remaining 5 regions showing greatest similarity with the *gag* and *pol* regions of murine leukemia viruses. Two of the regions were composites of more than one reading frame. In these cases, searches showed immediately adjacent HaSV regions in different frames matching immediately adjacent regions of viral sequences. This was interpreted to mean that a mutation in the HaSV sequence had split the original coding sequence between two or more reading frames. The sequences from the different reading frames were combined and searched against the database again, treating the two combined sequences as one.

Residues 1160 to 1420 of the VL30 showed the greatest similarity (110 residue identity out of 210) to the C-terminal region of the retroviral *pol* polyprotein, which is cleaved to yield an endonuclease. Confirmation of the amino acid sequences was found by searching for nucleic acid similarities. A 62% identity was found over a 625-base overlap with Moloney leukemia virus sequences. In addition, the nucleic acid alignments showed insertions or deletions corresponding in location to the frameshifts introduced into the peptide sequences. The similarity includes an (imperfect) copy of the C-X2-C-X4-H-X4-C motif, known to be conserved in these sequences.

Manly *et al.* conclude that the sequence comparisons suggest a distant evolutionary relationship between rat VL30 sequences and murine leukemia virus sequences. A relationship between the VL30 sequences and sequences of the retrovirus group including feline sarcoma virus and baboon endogenous virus are also suggested by the results of the database searches. Further, the likelihood that VL30 sequences directly code for an anoxic stress protein (lactate dehydrogenase k) is considered unlikely, since the searches failed to find any similarity with dehydrogenase sequences. However, their extensive similarity with endonuclease sequences suggests that they may code for a protein with that function instead.

"Structure and Function of the Yeast tRNA Ligase Gene". Shawn K. Westaway, Eric M. Phizicky, and John Abelson. ***Journal of Biological Chemistry 263*****: 3171-3176 (1988).**

Dr. Abelson's laboratory has pioneered in the study of tRNA splicing in yeast, and in this paper they describe the DNA sequence of the entire coding region of the *Saccharomyces cerevisae* tRNA ligase gene. tRNA ligase is one of two enzymes required for tRNA splicing in yeast. The substrates for splicing are a subset of tRNA precursors containing introns. There are no obvious conserved regions at the splice junctions (unlike the case with introns in mRNA precursors) and the only common

feature is location, which is one based removed from the 3'-end of the anticodon.

The tRNA ligase molecule itself is a single ~90-kDa polypeptide likely to contain the three separate activities required for tRNA splicing - namely phosphorylation of the 5' terminus of the 3' half-tRNA in the presence of ATP; opening of the 2',3'-cyclic phosphodiester bond of the 5' half-tRNA; and ligation of the two tRNA halves. In order to study the functional domains of this protein the gene was cloned from *S. cerevisae* and its DNA sequence determined. Westaway *et al.* cloned into M13 phage using four unique EcoRI restriction sites to subclone plasmid pUC12-RLG. Purified phage DNA templates were sequenced by the Sanger dideoxy method by first using the M13 primer and then by priming with synthetic oligonucleotides. The 4.2 kb EcoRI fragment containing the yeast tRNA ligase gene was sequenced by obtaining a restriction map comprised of the M13 and synthetic oligonucleotide primers. Sites used were: EcoRI (site A), HpaI, ScaI, BglII, HindIII, KpnI, EcoRV, XbaI, and EcoRI (site B). M13 sequencing primer was used to obtain initial sequencing information. Oligonucleotides corresponding to unique regions of nucleotide sequence were then used as primers in directing continued synthesis along the same strand of each clone. Thus the sequence of both strands of each restriction fragment was obtained. To confirm the sequence at restriction site junctions, separate overlapping clones which spanned the junctions were sequenced.

The sequences of each fragment were entered onto BIONET using the GENED sequence editor, and the entire tRNA ligase sequence assembled using the GEL program. Prior knowledge of the initiator codon allowed immediate recognition of the tRNA ligase open reading-frame. Two further open reading-frames, ORF1 and ORF2, were also discovered using the SEQ/TRANSLATE software on BIONET. The mature tRNA ligase molecule was found to be 827 amino acids long with a molecular mass of 95.4 kDa. It is a basic protein (as expected for one involved in tRNA metabolism). Analysis using the PEP program showed that greater than 10% of the amino acids are lysines, and the net charge of the protein is +12.5 (counting histidine as +0.5). Codon usage frequencies, calculated using SEQ, show that tRNA ligase uses a large percentage of rarely used codons, consistent with the hypothesis that less abundant proteins (tRNA ligase is present in approximately 400 copies/yeast cell) do not have the bias toward preferred codons seen in highly abundant proteins. Suspicions that tRNA ligase expression might be controlled by levels of intron-containing tRNAs were proven unfounded when the codon usage of tRNA ligase relative to codons in yeast which are specifically translated by tRNAs (whose precursors contain introns) were compared using the SEQ program.

Despite the similarity in mechanism of action between yeast tRNA ligase and that of T4 RNA ligase and T4 DNA ligase, no obvious sequence similarities could be detected between the yeast tRNA ligase gene product and any of the other T4 ligases when compared with the IFIND program. Further, a screening of the entire NBRF-PIR protein sequence database using both IFIND and XFASTP revealed no significant similarities between T4 ligase and any database sequences. Other data suggesting that tRNA ligase has clearly separable domains responsible for some or all of the different activities observed during tRNA splicing should allow more subtle subsequence similarities between tRNA ligase and other proteins of similar function to be discerned.

Westaway *et al.* summarise the properties of the tRNA ligase gene product as:1) being able to catalyze three different activities; 2) forming a splicing complex with endonuclease and tRNA precursors; and 3) being localized to a specific site at or near the inner nuclear membrane. Further studies of the gene and protein product will be neccessary to characterise how these different aspects

are embodied in a single polypeptide.

3.1.1.2 Scientific Consulting: BIONET User Support

The Service component of the BIONET Resource is supported by a group of three BIONET staff members as described below. The staff interacts with the community in a variety of ways, including direct support via telephone calls, electronic mail and terminal links with individual investigators. Support is also provided through staff participation at major meetings and trade shows, at trainings, and through participation in providing on-line and printed documentation for User Manuals, program descriptions and system procedures.

We currently have a full time BIONET Scientific Consultant (Dr. Karen Davis), a full-time Applications Analyst (Mr. Spencer Yeh), and a full time Senior Scientific Consultant (Dr. Sunil Maulik). The Scientific Consultant provides direct support (telephone and e-mail assistance) to the community 75% of her time on a rotating basis. The other 25% of her time is devoted to the development of the BIONET training program and other Service projects. Tha Applications Analyst devotes 25% of his time to user support and the rest of his time to contributed software and database development/maintenance tasks. The Senior Scientific Consultant oversees the electronic mail responses of the Consultant and Analyst, assists them in answering more complex questions, and spends the remainder of his time participating in Technological and Collaborative Research.

3.1.1.3 Service

The Service component of the BIONET Resource includes primarily Class I investigators and takes the form of answering questions by phone, by electronic mail, and by terminal links from investigators to staff. A survey of the monthly phone, mail, and terminal links for the past year shows the different uses of the BIONET Resource by the user community. The monthly inquiry rates broken down into several categories are given below.

Table 3-1: Summary of Monthly Rates of Inquiries

Category	Number of Inquiries	Percent of Total Inquiries
Programs and Databases	164	52
TOPS20 System	54	17
Administration	28	9
Electronic Mail	25	8
PC and PC Software	20	6
Telecommunications	14	4
File Transfers	14	4
TOTALS	319	100

The yearly total is 3813 queries, or about 15 per day (based on a 260- day work year). As can be seen from Table III-1, the largest number of inquiries--52%--concern the use of BIONET's programs and databases. This Programs-and-Databases category has been subdivided into additional scientific and program categories in Table III-2.

Table 3-2: Summary of Monthly Rates of Questions for Programs and Databases.

Category	Number of Inquiries	Percent of Total Inquiries
Database Searches and Databases	85	52
DNA and Protein Sequence Analysis	27	16
Sequence and Gel Data Entry and Manipulation	19	12
Experiment Planning and Analysis	6	4
Multi-Sequence Alignment	5	3
TOPS20 System Programs	13	8
Other	9	5
	----	----
TOTALS	164	100

As shown in Table III-2, just over half of these questions about Programs and Databases concern Databases and Database Searches. This undoubtedly reflects the convenient access to fast database-searching programs and recent versions of sequence databases on BIONET. As mentioned below, the FASTA-MAIL program, introduced this year to the BIONET system, has been especially popular because of its speed and ease of use.

Table III-2 also shows that the breakdown of the other categories of Programs-and-Databases questions is as follows: 16% on Sequence Analyses; 12% on Sequence and Gel Data Entry and Manipulation; and less than 10% on each of the other categories. Inquiries about Multi-Sequence Alignment mostly concerned the use of the IntelliGenetics' GENALIGN program and William Bains' contributed XMULTAN program. Questions about TOPS20 System Programs concerned the use of the FIND and XSEARCH programs to examine database-index files, and the use of the XGENPUB program to submit sequences to the GENBANK, EMBL, and PIR databases. Inquiries about

Experiment Planning and Analysis concerned the use of the SIZER, MAP, and CLONER programs. The category OTHER covers inquiries on the use of other contributed programs, such as Michael Zuker's BIOFLD, and other miscellaneous questions.

Returning to Table III-1, the second largest category--17%--of questions listed includes those concerning the TOPS20 operating system. These are separate from the the TOPS20 system programs listed in Table III-2. These questions concerned manipulating files and directories, using the text editors, and logging in to one's BIONET account.

The third largest category (9%) comprised BIONET Administrative questions which came directly to the consultants and were rerouted to the BIONET Administrator. The BIONET administration category consisted mostly of application requests, training session information, and manual requests. In addition to the calls listed here there are many calls directly to the BIONET Administrator which are not included in Table III-1.

The fourth largest category in Table III-1 is Electronic Mail at 8%. The Telecommunications category, at 4%, includes both inquiries concerning the procedures involved in connecting to the BIONET computer and the quality of the communications between remote users and BIONET. The remaining categories are self-explanatory.

The number of inquiries this year (3813) and the rate per day (15) are only 4% higher than last year, when there were 3700 queries, or about 14 per day. We believe that this constancy of rate reflects a balance between several factors: (1) fewer inquiries per user, due to new documentation written this year, but (2) counterbalanced by an increased number of users; and (3) a slightly greater increase in the number of new accounts. This year 270 new labs opened accounts on the system versus 238 new accounts last year, an increase of 13%. However, new documentation prepared this year includes more on-line examples of the uses of the programs in the analysis of a research project and a major revision of the *Introduction to BIONET* manual, which is sent to all new users. This documentation is described below.

Not only was the total number of inquiries approximately the same this year as last year, but also the relative ranking of the various categories of inquiries on the basis of percent of inquiries was similar. The main difference is the higher proportion this year of questions on databases and database searches.

All of the above data indicate that the consultant service is an extremely important component of the BIONET Resource. There are so many features available on BIONET that the presence of a trained expert can assist users in utilizing the resource efficiently and intelligently. In addition, the large number of questions pertaining to the databases and the database searching programs point to a major use of the Resource for large scale database searching and analyses.

3.1.1.4 PC/BIONET Communications - Distribution of the Resource

It has been clear from the beginning of operation of BIONET that the majority of the user community had access to personal computers, and that they all were looking for ways to use the PC's effectively in conjunction with BIONET. We have strongly supported this method of access, to the extent of maintaining a lending (and on-line) library of software and documentation for file transfer and terminal emulation programs. We have worked closely with the community in this way because

we recognize that distribution of the Resource would be required in order for the central DEC-2065 to support an ever-increasing number of users.

If the only use of PC's was as terminals and as a source or recipient of files transferred over the network, the net burden on BIONET would probably increase, rather than decrease. The availability of PC-based software for sequence analysis has provided scientists with a means for performing many simple analyses locally. When they need access to BIONET for a more complicated analysis, they merely log on and transfer any needed files to the DEC-2065. As this "style" of computation increase in popularity, the burden on the DEC-2065 will be reduced significantly.

To facilitate use of PC's, our Consultants provide information on file formats and the use of editors on BIONET to reformat sequence files uploaded from PC's. In addition, several PC molecular biology programs are distributed via the BIONET software lending library. These programs allow users to perform many routine analyses locally. The programs are described below under *Computer Software - Contributed Library.*

The IBM PC public domain version of BIONET's on-line EMACS editor has been furnished via the lending library. "MicroEMACS" is distributed free to users and has the virtues of producing ASCII text files compatible with the BIONET software and of utilizing essentially the same command set as the mainframe editor. As new versions become available BIONET has updated its lending library diskettes and announced the availability to users via the electronic bulletin board system. Use of MicroEMACS facilitates sequence entry on user PC's and reduces dependence on the mainframe.

BIONET has also sought to standardize file transfer protocols on the system by vigorously promoting the use of the public-domain Kermit software. New versions of Kermit have been added to the lending library as they have become available. This last year saw the inclusion of a new version of Kermit for the IBM-PC which supported Tektronix 4014 emulation and automatic login macros. The former feature allowed users to obtain graphic output of plasmid maps from the IntelliGenetics (IG) CLONER program and dot-matrix plots from the IG DDMATRIX program. The latter feature (login macros) automated the process of dialing in to the BIONET Resource over Telenet or Compuserve.

3.1.1.5 FASTA-MAIL program

On August 17, a new program was introduced on BIONET that has shortened the turnaround time on database similarity searches by a factor of 20, increased the interactive response on the BIONET DEC machine by a factor of 3, while at the same time maintaining excellent sensitivity to biologically significant similarities. This was accomplished by utilizing BIONET's in-house network capabilities to set up a remote database server on the new BIONET Sun 3/280 computer using the FASTA contributed software program from David Lipman and William Pearson. The FASTA-MAIL program was based on the earlier FASTP-MAIL project which allowed BIONET users to submit remote FASTP database searches of the NBRF/PIR database on a Sun 3/280. The FASTA-MAIL program incorporated the following enhancements: a simpler user interface, the ability to do nucleic acid database searches of GenBank, the additional capability to search SWISS-PROT, the ability to set the KTUP parameter, the use of the more sensitive FASTA algorithm instead of the older FASTP algorithm, and significantly expanded on-line documentation for using the program. Since its introduction on BIONET, the FASTA-MAIL program has been extremely well-received and is used 950 times per month, one of the most popular programs on BIONET.

Extensive work was required by the BIONET staff on the FASTA-MAIL project. The FASTA program itself had to be modified to accept IntelliGenetics' file format, to have a "brief output" option to limit the alignment output to the aligned region only, and to improve the clarity and labeling of all output. The "brief output" modification was necessary since without this modification, the program frequently created output files that were larger than the BIONET mailer could handle (256 Kbytes) in a single message. The actual user interface and remote mail server software was created by the BIONET systems programmers Eliot Lear and Rob Liebschutz. This included a user interface on the DEC to submit automatically a formatted mail message to the correct address on the Sun computer, authorization software, queueing software to provide separate protein and DNA search queues on the Sun, software for running the FASTA program non-interactively, and software for mailing the search results to the correct user account on the BIONET DEC machine. Finally, two separate on-line help topics were written to provide an explanation of the FASTA algorithm and to give a step-by-step guide to using the program and accessing the output results through the MM mail program on the DEC 2065.

The FASTA-MAIL program allows both nucleic acid searches and protein searches to be run with the same program by using different scoring matrices and program switches for the two types of searches. Since the FASTA program gains some of its speed through the use of a stripped version of the databases, FASTA formatted versions of the GenBank, NBRF/PIR, and SWISS-PROT databases were created. All together, supporting these three databases for FASTA searches currently requires an additional 30 Mbytes of disk space beyond the normal requirements for these databases. We are also planning to implement a FASTA-formatted version of the EMBL nucleic acid sequence Data Library.

The user interface for the FASTA-MAIL program is designed to be extremely easy to use. The user need only respond to four simple questions to set up the search; this only takes about fifteen seconds! Since the job is run remotely, the user can log off as soon as the job is submitted. The results are automatically deposited in his or her mail file when the search is completed. Average CPU times for the searches are 1 minute for full protein databank searches and 20 minutes for full GenBank searches. Actual turnaround times are determined by the number of jobs waiting to execute in the queue. Typical turnaround times are about 5-20 minutes for full protein databank searches and 1-4 hours for full GenBank searches. Since these CPU intensive searches would otherwise have been run directly on the BIONET DEC machine, the FASTA-MAIL program has had the positive effect of dramatically increasing the responsiveness of the DEC machine. Typical mid-day user load levels on the DEC-20 were in the range of 10 - 15 before FASTA-MAIL was implemented. Now the average is 3 - 4. This means that the BIONET DEC-20 machine appears to be running about 3 to 5 times faster to the typical user.

Finally FASTA-MAIL, which although extremely fast, is still highly sensitive to biological similarities. By using a PAM250 scoring matrix for proteins, functionally similar amino acids at corresponding positions increase the score of an alignment, and by using a joining penalty in the database scan step, sequences with insertion gaps in significant regions are still kept for optimal alignment at a later stage. This last feature is the major improvement of the FASTA algorithm over the FASTP and FASTN algorithms which sometimes dropped significant matches from consideration because of the presence of small gaps. BIONET would like to thank the authors of FASTA, David Lipman of the NIH and William Pearson of the Univ. of Virginia, for allowing the BIONET